HAI Stanford University
Human-Centered
Artificial Intelligence

POLICY BRIEF
JUNE 2022

SPECIAL HEALTHCARE SERIES

Toward Stronger FDA Approval Standards for AI Medical Devices

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou

AS THE DEVELOPMENT AND ADOPTION OF ARTIFICIAL INTELLIGENCE-ENABLED HEALTHCARE TOOLS continue to accelerate, regulators and researchers are beginning to confront oversight concerns in the clinical evaluation process that could yield negative consequences on patient health if left unchecked. Since January 2015, the United States Food and Drug Administration (FDA) has evaluated and granted clearance for over 100 AI-based medical devices using a fairly rudimentary evaluation process that is in dire need of improvement as these evaluations have not been adapted to address the unique concerns surrounding AI. In fact, the FDA itself recently called for improving the quality of the evaluation data, increasing trust and transparency between developers and users, monitoring algorithmic performance and bias on the intended population, and testing with clinicians in the loop. Although academics are starting to develop new reporting guidelines for clinical trials, there is currently a lack of established best practices for evaluating commercially available AI medical devices to ensure their reliability and safety.

KEY TAKEAWAYS

- We analyzed public records for all 130 FDA-approved medical AI devices between January 2015 and December 2020 and observed significant variety and limitations in test-data rigor and what developers considered appropriate clinical evaluation.

- When we performed an analysis of a well-established diagnostic task (pneumothorax, or collapsed lung) using three sets of training data, the level of error exhibited between white and Black patients increased dramatically. In one instance this meant that the drop in the rate of accuracy grew by a factor of 1.8 times and 4.5 times, respectively.

- To minimize risks for patient harm and disparate treatment, policymakers should set the standard of multi-site evaluations, encourage greater comparison between standard of care without AI-enabled tools with a potential use of AI tools, and mandate post-market surveillance of AI devices.

In the paper titled "How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals," we examined the evaluation process performed on 130 FDA-approved AI medical devices between January 2015 and December 2020. The shortcomings were significant: 97% performed only retrospective evaluations that are much less credible; 72% did not publicly report whether the algorithm was tested on more than one site; and 45% didn't report basics, like sample size. We show performance degradation—and potential demographic bias—when algorithms are tested on only a single site with a model designed to detect collapsed lungs in chest X-rays.

The findings from our research ultimately led us to the following three policy recommendations:

1. Ensure future FDA-approved AI devices undergo multi-site evaluations.
2. Encourage more prospective studies—i.e., those in which the test data is collected and evaluated concurrently with device deployment—that include a comparison to current standard of care without AI.
3. Mandate post-market surveillance of medical AI devices to better understand some of the unintended outcomes and biases not detected in the evaluation process.

Introduction

The number of FDA approvals for AI devices has increased rapidly in the past five years, with over 75 percent of approvals coming in the past two years and over 50 percent coming in the past year. The FDA publicly releases a summary document for each device containing information about the clinical evaluation process, such as the number of patients enrolled, the number of evaluation sites, indications for use, whether the test data was collected and evaluated concurrently with device deployment (what is termed a prospective evaluation) or if the test set was collected before device deployment (a retrospective evaluation), and performance data.

We created a comprehensive database containing the aforementioned details from the summary documents of over 100 FDA-approved AI medical devices to better understand if the FDA is meeting its goals of enhancing test-data quality, improving trust and transparency with users, monitoring algorithmic performance, and testing with clinicians in the loop.

We created a comprehensive database containing the aforementioned details from the summary documents of over 100 FDA-approved AI medical devices to better understand if the FDA is meeting its goals of enhancing test-data quality, improving trust and transparency with users, monitoring algorithmic performance, and testing with clinicians in the loop. We also assigned each device a risk level from 1 to 4, with 1 being the lowest risk and 4 being the highest risk, according to an FDA proposal.

Research Outcomes

Several key findings emerge from this analysis. First, most of the evaluations (126 of 130) were performed as retrospective studies only and none of the 54 high-risk devices were evaluated by prospective studies. The difference between prospective and retrospective device evaluation is significant because prospective studies provide a more comprehensive understanding of how an AI model affects clinical practice, especially since human-computer interactions can deviate from a model's original purpose, and are less likely to be gamed. For example, a prospective study may reveal that clinicians are misusing an AI decision support device for primary diagnosis.

Another surprising finding was the lack of specificity regarding the reported geographic locations where the devices were evaluated. We found that 93 of the approved devices lacked public information on whether a multi-site assessment was part of the evaluation. Further, of the 41 devices that did include information about the number of assessment sites, four devices were evaluated at one site and eight devices were evaluated at two sites. Knowing whether a device underwent multi-site evaluations is critical for understanding model reliability and bias as it accounts for variations in patient demographics or technician standards. Only 17 device studies mentioned demographic subgroup performance such as gender and race in their evaluations.

> *Knowing whether a device underwent multi-site evaluations is critical for understanding model reliability and bias as it accounts for variations in patient demographics or technician standards.*

Even standard study details like sample size are not consistently reported. The published reports for 59 devices did not include the sample size. And the median evaluation sample size was small (300) for instances where it was reported.

To understand the challenges of generalizing a model beyond a few hospital sites, we conducted a case study to measure the performance of a pneumothorax (commonly referred to as collapsed lung) detection model across three hospital sites: the National Institutes of Health (NIH) Clinical Center in Bethesda, MD; Stanford Health Care in Palo Alto, CA; and Beth Israel Deaconess Medical Center in Boston, MA. Using a top-performing deep-learning model for classifying chest conditions, we trained three separate models on the data from patients at each of the three sites and then evaluated the prediction models against one another.

Specifically, the model trained on NIH data achieved good performance on independent NIH test patients but performed much worse on the Beth Israel Deaconess and Stanford Health Care test patients. Across all three datasets, we found substantial drop-offs in prediction model accuracy when the models were evaluated at a different site. We attributed this level of variation to the differences in patient demographics across sites. For example, for the three prediction models evaluated on the Beth Israel Deaconess test patients, we found a serious discrepancy in the error between white and Black patients, which reduced accuracy by a factor of 1.8 and by 4.5 times, respectively.[1] Assessments with only a few sites—a prevalent practice according to our dataset—is a poor metric for gauging a model's performance on larger populations.

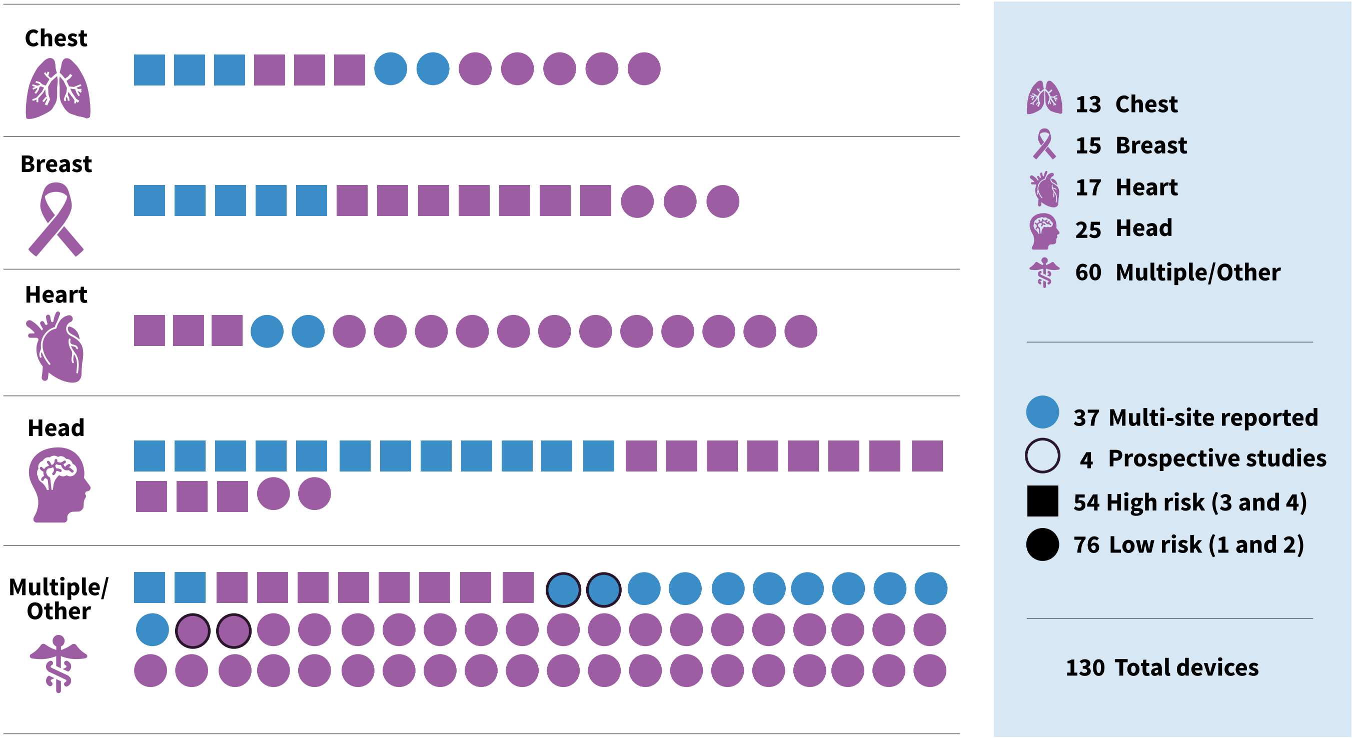
> *Our database suggests that a substantial proportion of approved devices may have been evaluated at a small number of sites, limiting both geographic and population diversity.*

Policy Discussion

Policymakers should prioritize preventing adverse medical treatment as a result of AI. The pneumothorax analysis shows how crucial multi-site evaluations are for understanding a model's bias and reliability. Multi-site evaluations help account for variations in the equipment used, technician standards, image-storage formats, demographic makeup, and disease prevalence. Our database suggests that a substantial proportion of approved devices may have been evaluated at a small number of sites, limiting both geographic and population diversity. Further, full details about evaluation sites allow clinicians, researchers, and patients to make informed judgments about the reliability of the algorithm.

Policymakers have the power to ameliorate these shortcomings. First, policymakers can set the standard that medical AI devices must undergo multi-site evaluations to ensure devices perform well across the country, and summary documents must include information regarding site evaluations. Second, they can encourage prospective studies (those that are performed when the test data is collected and evaluated concurrently with device deployment) to reveal whether practitioners are using a device for its intended use. These prospective study results should be compared to the results of current methods doctors perform without AI systems to accurately capture true clinical outcomes. Third, policymakers can mandate post-market surveillance of AI devices so researchers can better understand and measure the unintended outcomes and biases that are not detected in prospective, multi-site trials.

[1] Performance disparity between white and Black patients increased from 0.024 AUC with the BIDMC-trained model to 0.043 AUC and 0.109 AUC with the other two models.



Breakdown of 130 FDA-approved medical AI devices analyzed: Devices are categorized by risk level (square, high risk; circle, low risk). Blue indicates that a multi-site evaluation was reported; otherwise, symbols are purple. Black outline indicates a prospective study (key, right margin). Numbers in key correspond to the number of devices with each characteristic.

The original article, "**How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals**" can be accessed at https://www.nature.com/articles/s41591-021-01312-x.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Stanford University
Human-Centered
Artificial Intelligence

Eric Wu is a PhD student in the department of electrical engineering at Stanford University.

Kevin Wu is a PhD student in the department of biomedical informatics at Stanford University.

Roxana Daneshjou is a clinical scholar in dermatology and a postdoctoral research fellow in biomedical data sciences at Stanford University.

David Ouyang is a cardiologist and researcher in the department of cardiology and the division of artificial intelligence in medicine at Cedars-Sinai Medical Center.

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, professor of political science, a senior fellow at the Stanford Institute for Economic Policy Research, the director of the Regulation, Evaluation, and Governance Lab (RegLab), and an associate director at the Stanford Institute for Human-Centered Artificial Intelligence at Stanford University.

James Zou is assistant professor of biomedical data science and, by courtesy, of computer science and of electrical engineering at Stanford University.

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**